39. (Amended) A [gene, or a fragment thereof,] purified polynucleotide comprising DNA having at least [50%] 60% identity with SEQUENCE ID NO 4 or SEQUENCE ID NO 5.

40. (Amended) The method of claim 1 wherein the presence of said target [BS124] polynucleotide in said test sample is indicative of breast disease.

II. REMARKS

Claims 1-44 are pending. Claims 17-29, 31, 32, 34, 36, 37, 43 and 44 have been withdrawn pursuant to a restriction requirement. Claims 1-16, 30, 33, 35 and 38-42 stand variously rejected under 35 U.S.C. §§ 112, first and second paragraphs, 102 and 103.

By amendment herein claims 1-3, 6, 10, 11, 14, 15, 30, 33 and 38-40 have been amended. Amendment of these claims is not intended to be an acquiescence in the Office's assessment of those claims, and Applicants expressly reserve the right to bring the subject matter of the original claims again in a subsequent, related application.

Basis for the amendments and newly added claims can be found as follows. Support for the amendment to claims 1, 10, 11, 14, 30, 33 and 39 can be found, for example, in the original claims and sequence listing on page 14, line 33 ("preferably about 60% identity) and on page 4, line 24 ("diagnostic"). Basis for the amendments to claims 3, 6 and 10 can be found, for example, at page 13, line 2 and page 20, line 27 ("indicative of breast tissue disease"); at page 14, line 15 ("10 nucleotides"); and on page 14, line 34 ("90% identity"). Thus, no new matter has been added by way of this amendment and the entry thereof is respectfully requested.

Rejection of Claims 1-16, 30, 33, 35, 38-42 under 35 U.S.C. §112, First Paragraph

The Examiner has rejected claims 1-16, 30, 33, 35, 38-42 under 35 U.S.C. § 112, first paragraph, asserting that the specification does not enable polynucleotides having "at least 50% identity with" the recited Sequence Identification Numbers.

Applicants traverse the rejection and its supporting remarks.

By law, a patent application is presumptively enabled when filed. *In re Marzocchi*, 169 USPQ 367, 369 (CCPA 1971). In other words, without reason to doubt the truth of the statements made in the patent application, the application must be considered enabling. see, *In re Wright*, 27 USPQ2d 1510, 1513 (Fed. Cir. 1993) and *In re Marzocchi, supra*. Thus, the burden is on the Office to establish why the claimed invention is not enabled by the specification.

Applicants disagree with the Examiner's assessment of the level of enabling disclosure in the present applicant in regard to "percent identity." The applicants discuss the use of available programs for calculating identity or similarity between sequences in the specification (e.g., page 12, lines 16-35). In addition, Applicants submit the software manual to the Wisconsin Sequence Analysis program, version 8, publicly available from Genetics Computer Group, Madison, WI (Exhibit A), as discussed on page 12, line 31 of the specification. The manual provides the algorithm, parameters, parameter values and other information necessary to, accurately and consistently, calculate percent identity. This manual indicates on pages 5-21, *inter alia*, that the software used the local homology algorithm of Smith and Waterman (Advances in Applied Mathematics 2:482-489 (1981). Thus, Applicants submit that use of default parameters in such programs is routine and well within the abilities of one having ordinary skill in the art -- this is the manner in which the Examiner has searched the database for sequences that may correspond to the claimed sequences.

Rejection of Claims 1-16, 30, 33, 35, 38-42 under 35 U.S.C. §112, Second Paragraph

The Examiner has rejected claims 1-16, 30, 33, 35, 38-42 under 35 U.S.C. § 112, second paragraph, asserting that the claims are indefinite for failing to particularly point out and distinctly claim the subject matter which the applicant regards as the invention. The Examiner has asserted the following specific deficiencies in the claims.

A. "Percent identity"

The Examiner asserts that recitation of "% identity" is vague and indefinite. Applicants disagree with the Examiner's assessment of the level of enabling disclosure in the present applicant in regard to "percent identity." The applicants discuss the use of available programs for calculating identity or similarity between sequences in the specification (e.g., page 12, line 16-35). As noted above in addressing the section 112, first paragraph rejection, Applicants submit that use of default parameters in such programs is routine and well within the abilities of one having ordinary skill in the art -- this is the manner in which the Examiner has searched the database for sequences that may correspond to the claimed sequences. (see, also Exhibit A, attached hereto). Further, at the AIPLA meeting in Crystal City, Fall of 1999, Examiner John Doll stated that the USPTO policy toward claims reciting percent identity has changed and that Examiners will no longer be rejecting percent identity claims under 35 U.S.C. §112, second paragraph.

B. "Complement Thereof"

The Examiner has also objected to the term "complement thereof" used in various claims.

Applicants traverse. The standard is that the "definiteness of the language must be analyzed...in light of the teachings of the prior art and of the particular application disclosure as it would be interpreted by one possessing the ordinary level of skill in the pertinent art." *In re Moore, supra.* A claim which is clear to one ordinarily skilled in the art when read in light of the specification, does not fail for indefiniteness. *Slimfold Mfg. Co. v. Kinkead Indus., Inc., supra.*

Nothing in the recitation of "complement thereof" would be unclear to a skilled artisan (see, *e.g.,* page 13, line 32 and page 14, line 16). The complement of a particular nucleic acid sequence can be readily determined by one possessing ordinary skill in the art. Such a skilled artisan would readily understand and interpret what is meant by "and complements thereof." As defined for example in the Dictionary of Biotechnology,

(1986, 2nd ed.), complementary sequences are "two sequences of nucleotides that are capable of base pairing throughout their length." (copy attached hereto as Exhibit B). Thus, in view of the teachings of the specification and teachings of the prior art, the recitation of "and complements thereof" is clear to one skilled in the art and Applicants respectfully request withdrawal of this rejection.

C. "Fragments and Complements Thereof"

The Examiner asserts that recitation of "and fragments and complements thereof" is vague and indefinite.

Applicants have amended the claim to clarify the Markush group thereby obviating this rejection.

D. "Derived From"

The Examiner asserts that recitation of "derived from" is vague and indefinite in claims 11-14 and 16.

The term "derived from" is defined in the specification (see, for example, page 13, lines 26-35). However, in order to facilitate prosecution, Applicants have removed this language from the claims.

E. "BS124"

The Examiner has asserted that the claims are vague and indefinite in the recitation of "BS124."

The term "BS124" is described extensively throughout the specification (see, for example, pages 4-10; and Example 1). However, in order to facilitate prosecution applicants have removed this language from the claims.

F. "Hybridizes Selectively"

The Examiner asserts that recitation of "hybridizes selectively" is vague and indefinite in claim 11.

The term selectively "hybridizes selectively" has an art recognized meaning to one of ordinary skill in the art. Examples of types of selective hybridization are described throughout the specification (see, for example, pages 27-32), as well as, specific examples of using the polynucleotide sequences of the present invention in methods involving selective hybridization (see, for example, pages Examples 4-9). Accordingly, the applicants submit that the language "hybridizes selectively" would be clear to one of ordinary skill in the art. However, in order to facilitate prosecution applicants have removed this language from the claim and the rejection of claim 11 under 35 U.S.C. §112, second paragraph, should be withdrawn.

G. Claims 15 and 16

Claims 15 and 16 are alleged to be unclear indefinite in the recitation of the phrase "comprising a nucleic acid sequence that includes an open reading frame derived from BS124 operably linked to a control sequence compatible with a desired host."

Although Applicants submit that the claims would be sufficiently clear to one skilled in the art, they thank the Examiner for her suggested language and have incorporated that language by amendment herein. Accordingly, the rejection has been obviated.

In view of the foregoing remarks and amendments, the Applicants submit that the pending claims comply with the requirements of 35 U.S.C. §112, second paragraph, and the rejection of the claims should be withdrawn.

Rejection of Claim 11 Under 35 U.S.C. §102(b)

The Examiner has rejected claim 11 under 35 U.S.C. §102(b) asserting that the claim is anticipated by GenBank Accession Number AC002098.

As a threshold matter, Applicants request clarification of the rejection which states that AC002098 is "99.2% identical to base pairs 5815-5694 of SEQ ID NO:1." (Office Action, page 9, paragraph 4). Applicants note that SEQ ID NO:1 is only 236 base

pairs in length. Further, even assuming that the numbering in the Office Action refers to the query sequence (AC002098), Applicants note that the "query match" is only 51%, not 99.2% as stated in the Action. Thus, clarification of this rejection is respectfully requested.

In any event, AC002098 does not anticipate pending claim 11. For prior art to anticipate under 35 U.S.C. 102 it has to meet every element of the claimed invention: such a determination is one of fact. *Hybritech Inc. v. Monoclonal Antibodies*, 802 F.2d at 1367, 231 USPQ 81 (Fed. Cir. 1986).

As amended herein, claim 11 is drawn to the specific sequences of SEQ ID Nos: 1, 2, 4, 5 and complements thereof. SEQ ID NO:1 is 236 base pairs in length. As noted above, AC002098 exhibits, at best, sequence similarly to a fragment of SEQ ID NO:1 extending from base pair 115 through 236. Thus, when aligned over its entire length, the sequence similarity falls well below the claimed 50% limit. In sum, the prior art cited by the Examiner fails to teach a polynucleotide having (i) at least 50% identity to the polynucleotide of SEQ ID Nos: 1, 2 or complements thereof or (ii) at least 60% identity to the polynucleotide of SEQ ID Nos: 4, 5 or complements thereof.

In view of the above amendments and arguments, the cited reference sequence cannot be said to teach all the elements of the present invention. Accordingly, there is no support for the pending claims being anticipated by the cited prior art under 35 U.S.C. §102(b) and withdrawal of the rejection is respectfully requested.

Rejection of Claim 11 Under 35 U.S.C. §102(a)

The Examiner has rejected claim 11 under 35 U.S.C. §102(a) asserting that the claim is anticipated by GenBank Accession Number AC002320.

Applicants again request clarification of the rejection which states that AC002320 is "99.2% identical to base pairs 48493-48372 of SEQ ID NO:1." (Office Action, page 9, paragraph 5). SEQ ID NO:1 is only 236 base pairs in length. Further, even assuming that the numbering refers to the query sequence (AC002320), Applicants note that the "query match" is only 51%. Therefore, clarification of the rejection is requested.

For the same reasons as detailed above with regard to GenBank Accession No. AC002098, the precisely claimed sequences recited in claim 11 are not anticipated by GenBank Accession No. AC002320. In particular, AC002320 exhibits only 51% identity to the fragment of SEQ ID NO:1 extending from base pair 115 through 236. The percent identity is much lower when the precisely claimed sequence is aligned. Thus, the prior art cited by the Examiner fails to teach a polynucleotide having (i) at least 50% identity to the polynucleotide of SEQ ID Nos: 1, 2 or complements thereof or (ii) at least 60% identity to the polynucleotide of SEQ ID Nos: 4, 5 or complements thereof. Accordingly, withdrawal of this rejection is requested.

Rejection of Claims 11-16, 30, 33, 35, 38 and 39 Under 35 U.S.C. §102(a)

The Examiner has rejected claims 11-16, 30, 33, 35, 38 and 39 under 35 U.S.C. §102(a) as allegedly anticipated by GenBank Accession No. AI1251747. In particular, AI1251747 is alleged to be 98.5% identical to base pairs 114-245 of SEQ ID NO:2; 99.7% identical to SEQ ID NO:3; 99.1% identical to base pairs 226-692 of SEQ ID NO:4 and 99.1% identical to base pairs 226-692 of SEQ ID NO:5.

Applicants note that GenBank Accession No. AI1251747 is not prior art against the pending application. In particular, the date of availability of the GenBank sequence used for searching was November 5, 1998, which is after both Applicants' filing date and priority date. As indicated in Exhibit C attached hereto, the date accorded to the particular sequence is the date listed to the right of the heading "PRI", not the date of journal submission. Indeed, because revisions and/or updates to GenBank sequences can be made at any time (see, Exhibit C, attached hereto), the date of journal submission may represent a completely different sequence than the one actually used for searching purposes. On this basis alone, the rejection is improper and should be withdrawn.

Even assuming, for the sake of argument only, that this sequence had been available prior to Applicants' invention thereof, this rejection is improper for reasons detailed above, namely that the actual percent identity (*i.e.*, "query match") is much lower than indicated in the Office Action. Indeed, when properly aligned over the entire length

of SEQ ID Nos: 1, 2, 4 and 5 (SEQ ID NO:3 has been canceled from the claims), the GenBank sequence does not approach the claimed percent identities. With regard to claim 38, there is no teaching in the reference to pick a small fragment of the prior art sequence and to translate this fragment into a polypeptide. Moreover, the fragment of AI1251747 chosen by the Office itself encodes only a portion of SEQ ID NO:22 (i.e. 110 amino acids out of 170 amino acids). Withdrawal of this rejection is, therefore, requested.

Rejection of Claims 11-16, 30, 33, 35, 38 and 39 Under 35 U.S.C. §102(b)

The Examiner has rejected claims 11-16, 30, 33, 35, 38 and 39 under 35 U.S.C. §102(b) as allegedly anticipated by GenBank Accession No AA460323 or, alternatively, GenBank Accession Number AA460385. (Office Action, pages 11-12, paragraphs 7 and 8). AA460323 is alleged to be identical to nucleotides 354-690 of SEQ ID NO:4 and AA460385 is alleged to 99.8% identical to base pairs 277-688 of SEQ ID NO:4 and 99.8% identical to nucleotides 277-688 of SEQ ID NO:5.

Neither AA460323 or AA60385 anticipate the subject matter of claims 11-16, 30, 33, 35, 38 and 39. Applying the same reasoning as described above, Applicants note that AA460323 and AA460385 exhibit sequence similarity only to specific fragments of the claimed sequences. Thus, the reference sequences do not teach the precisely claimed polynucleotides and polypeptides of these claims. Accordingly, withdrawal of this rejection is respectfully requested.

Rejection of Claims 11-16, 30, 33, 35 and 38-40 Under 35 U.S.C. §102(b)

The Examiner has rejected claims 11-16, 30, 33, 35 and 38-40 under 35 U.S.C. §102(b) as allegedly GenBank Accession Number AI143970. (Office Action, paragraph 9).

Applicants traverse this rejection. GenBank Accession No. AI143970 is not prior art against the pending application. In particular, the date of availability of the GenBank

sequence used for searching was November 5, 1998 after both Applicants' filing date and priority date. On this basis alone, the rejection is improper and should be withdrawn.

Further, for the same reasons as detailed above, namely the fact that AI143970 does not teach the precisely claimed polynucleotides and polypeptides of the pending claims, withdrawal of this rejection is respectfully requested.

Rejections of the Claims 1-9 Under 35 U.S.C. §103

The Examiner has rejected claims 1-9 under 35 U.S.C. §103(a) as being unpatentable over Nangaku et al in view of GenBank Accession Nos: AC002098, AC002330, AI1251747; AA460323, AA460385 or AI143970.

Applicants traverse this rejection.

It is axiomatic that obviousness cannot be established by combining teachings in the prior art absent some teaching or suggestion in the prior art that the combination be made. E.g., *In re Stence,* 828 F. 2d 751, 4 USPQ2d 1071 (Fed. Cir. 1987); *In re Newell*, 891 F. 2d 899, 13 USPQ2d 1248 (Fed Cir 1989). In particular, the fact that references can be combined does not make the combination obvious unless the prior art also contains something to suggest the desirability of that combination. *In re Sernaker*, 702 F.2d 989, 217 USPQ 1 (Fed,. Cir. 1983). The PTO has the burden of establishing a case of *prima facie* obviousness, and can meet this burden "only by showing some objective teaching in the prior art or that knowledge generally available to one of ordinary skill in the art would lead that individual to combine the relevant teachings of the references." *In re Fine*, 837 F.2d 1071, 5 USPQd2 1596 (Fed. Cir. 1988). No such objective teaching has been presented.

With regard to claim 1, as noted above, there is no teaching or suggestion within the references to arrive at the precisely claimed polynucleotides. None of the references disclose the precise sequences recited in claim 1. Thus, combining these distinguishable sequences with references teaching general methods of searching EST databases (Nangaku) does not establish a *prima facie* case of obviousness. Moreover, there is no guidance concerning the selection of the cited sequences from among the millions of

possible sequences available in the database (i.e., GENBANK or EMBL). Thus, the combination of references cited does not render claims 1 and 2, as amended, unpatentable.

Further, amended claims 3-9 each recite a limitation similar to the following: "a method of detecting the presence of a target polynucleotide indicative of breast tissue disease." None of the references singly or in combination teach that detection of the polynucleotides of the present invention may be indicative of breast tissue disease.

Accordingly, because the elements of the claimed invention are not taught by the cited references, the applicants submit that the rejections under 35 U.S.C. §103 should be withdrawn.

Rejections of the Claims 10 and 35 Under 35 U.S.C. §103

The Examiner has rejected claims 10 and 35 under 35 U.S.C. §103(a) as being unpatentable over Nangaku et al in view of GenBank Accession Nos: AC002098, AC002320, AI1251747; AA460323, AA460385 or AI143970 and in further view of Cohen (U.S. Patent No. 5,939,265).

Applicants traverse this rejection.

Claims 10 and 35 are directed to test kits useful for detecting a target polynucleotide indicative of breast tissue disease. As noted above, there is no suggestion within the references to arrive at the precisely claimed polynucleotides or, moreover, that detection of the polynucleotides of the present invention may be indicative of breast tissue disease. Thus, the combination of references cited does not render claims 10 and 35, as amended, unpatentable and Applicants respectfully request that this rejection be withdrawn.

III. CONCLUSION

Applicants respectfully submit that the claims comply with the requirements of 35 U.S.C. §112 and define an invention that is patentable over the art. Accordingly, a Notice of Allowance is believed in order and is respectfully requested.

If the Examiner notes any further matters which the Examiner believes may be expedited by a telephone interview, the Examiner is requested to contact the undersigned.

Please direct all further communications in this application to:

Mimi C. Goller, Esq.
Abbott Laboratories
D-377/AP6D-2
100 Abbott Park Road
Abbott Park, IL 60064-3500
Telephone: (847) 935-7550
Facsimile: (847) 938-2623.

Respectfully submitted,

Date: March 1st, 2000 By: D Pasternak

Dahna S. Pasternak
Registration No. 41,411
Attorney for Applicants

ABBOTT LABORATORIES
D-377/AP6D-2, 100 Abbott Park Road
Abbott Park, IL 60064-3500
Telephone: (847) 935-7550
Facsimile: (847) 938-2623

EXHIBIT A

FUNCTION

BestFit makes an optimal alignment of the best segment of similarity between two sequences. Optimal alignments are found by inserting gaps to maximize the number of matches using the *local homology* algorithm of Smith and Waterman.

DESCRIPTION

BestFit inserts gaps to obtain the optimal alignment of the best region of similarity between two sequences, and then displays the alignment in a format similar to the output from Gap. The sequences can be of very different lengths and have only a small segment of similarity between them. You could take a short RNA sequence, for example, and run it against a whole mitochondrial genome.

SEARCHING FOR SIMILARITY

BestFit is the most powerful method in the Wisconsin Sequence Analysis Package™ for identifying the best region of similarity between two sequences whose relationship is unknown.

EXAMPLE

The sequence gamma.seq contains an Alu family sequence somewhere in the first 500 bases. alu.seq contains a generic human Alu family repeat. The two sequences are aligned and the best segment of similarity is found with BestFit.

```
% bestfit

 BESTFIT of what sequence 1 ?  gamma.seq

                  Begin (* 1 *) ?
                End (* 11375 *) ?  500
              Reverse (* No *) ?

 to what sequence 2 (* gamma.seq *) ?  alu.seq

                  Begin (* 1 *) ?
                End (*   207 *) ?
              Reverse (* No *) ?

 What is the gap creation penalty (* 5.00 *) ?

 What is the gap extension penalty (* 0.30 *) ?

 What should I call the paired output display file (* gamma.pair *)

 Aligning ..........-..

          Gaps:      3
       Quality: 129.3
 Quality Ratio: 0.625
  % Similarity: 84.466
        Length:   209

%
```

OUTPUT

Here is the output file. Notice how BestFit finds and displays only the best segments of similarity:

```
 BESTFIT of: gamma.seq  check: 6474  from: 1  to: 500

Human fetal beta globins G and A gamma
from Shen, Slightom and Smithies,  Cell 26; 191-203.
Analyzed by Smithies et al. Cell 26; 345-353.

 to: alu.seq  check: 4238  from: 1  to: 207

HSREP2 from the EMBL data library
Human Alu repetitive sequence located near the insulin gene
Dhruva D.R., Shenk T., Subramanian K.N.; "Integration in vivo into
Simian virus 40 DNA of a sequence that resembles a certain family of
genomic interspersed repeated sequences"; Proc. Natl. Acad. Sci. USA
77:4514-4518(1980).  . . .

 Symbol comparison table: Gencoredisk:[Gcgcore.Data.Rundata]Swgapdna.Cmp
 CompCheck: 5234

        Gap Weight:  5.000      Average Match:  1.000
     Length Weight:  0.300   Average Mismatch: -0.900

           Quality:  129.3             Length:    209
             Ratio:  0.625               Gaps:      3
 Percent Similarity: 84.466   Percent Identity: 84.466

 gamma.seq x alu.seq       June 20, 1994 15:15  ..

     137 AGACCAACCTGGCCAACATGGTGAAATCCCATCTCTAC.AAAAATACAAA 185
         |||||| ||||||||||||||||||||  ||||||||||  ||||||||||
       1 AGACCAGCCTGGCCAACATGGTGAAACTCCATCTCTACTGAAAATACAAA 50

     186 AATTAGACAGGCATGATGGCAAGTGCCTGTAATCCCAGCTACTTGGGAGG 235
         |||||| |||||||| ||   |||||| ||||||||||||| |||||
      51 AATTAGCCAGGCATGGTGATGCGTGCCTGGAATCCCAGCTACTTAGGAGG 100

     236 CTGAGGAAGGAGAATTGCTTGAACCTGGAAGGCAGGAGTTGCAGTGAGCC 285
         |||||  || |||||__||| ||||  | |||  |_ ||||||||||||
     101 CTGAGACAGAAGAATCCCTTAAACCAAG.AGGTGGAGGTTGCAGTGAGCC 149

     286 GAGATCATACCACTGCACTCCAGCCTGGGTGACAGAACAAGACTCTGTCT 335
         ||||||_|| |_|||||||||||||||_|||||||||_|_||||||_|||
     150 GAGATCGCACGGCTGCACTCCAGCCT.GGTGACAGAGCGAGACTCCATCT 198

     336 CAAAAAAAA 344
         | | . .
     199 CAAAAAAAA 207
```

RELATED PROGRAMS

When you want an alignment that covers the whole length of both sequences, use Gap. When you are trying to find only the best segment of similarity between two sequences, use BestFit. PileUp creates a multiple sequence alignment of a group of related sequences, aligning the whole length of all sequences. DotPlot displays the entire surface of comparison for a comparison of two sequences. GapShow displays the pattern of differences between two aligned sequences. PlotSimilarity plots the average similarity of two or more aligned sequences at each position in the alignment. Pretty displays alignments of several sequences. LineUp is an editor for editing multiple sequence alignments. CompTable helps generate scoring matrices for peptide comparison.

ALGORITHM

BestFit uses the *local homology* algorithm of Smith and Waterman (Advances in Applied Mathematics 2; 482-489 (1981)) to find the best segment of similarity between two sequences. BestFit reads a scoring matrix that contains values for every possible GCG symbol match (see the LOCAL DATA FILES topic below). The program uses these values to construct a path matrix that represents the entire surface of comparison with a score at every position for the best possible alignment to that point. The *quality* score for the best alignment to any point is equal to the sum of the scoring matrix values of the matches in that alignment, less the gap creation penalty times the number of gaps in that alignment, less the gap extension penalty times the total length of all gaps in that alignment. The gap creation and gap extension penalties are set by you. If the best path to any point has a negative value, a zero is put in that position.

After the path matrix is complete, the highest value on the surface of comparison represents the end of the best region of similarity between the sequences. The best path from this highest value backwards to the point where the values revert to zero is the alignment shown by BestFit. This alignment is the best segment of similarity between the two sequences.

For nucleic acids, the default scoring matrix has a *match* value of 1.0 for each identical symbol comparison and -0.90 for each non-identical comparison (not considering nucleotide ambiguity symbols for this example). The *quality* score for a nucleic acid alignment can, therefore, be determined using the following equation:

```
Quality = 1.0 x TotalMatches +  -0.90 x TotalMismatches
                  - (GapCreationPenalty x GapNumber)
          - (GapExtensionPenalty x TotalLengthOfGaps)
```

The *quality* score for a protein alignment is calculated in a similar manner. However, while the default nucleic acid scoring matrix has a single value for all non-identical comparisons, the default protein scoring matrix has different values for the various non-identical amino acid comparisons. The *quality* score for a protein alignment can therefore be determined using the following equation (where $Total_{AA}$ is the total number of A-A (Ala-Ala) matches in the alignment, $CmpVal_{AA}$ is the value for an A-A comparison in the scoring matrix, $Total_{AB}$ is the total number of A-B (Ala-Asx) matches in the alignment, $CmpVal_{AB}$ is the value for an A-B comparison in the scoring matrix, ...) :

$$
\begin{aligned}
Quality = \; & CmpVal_{AA} \times Total_{AA} \\
& + CmpVal_{AB} \times Total_{AB} \\
& - CmpVal_{AC} \times Total_{AC} \\
& \vdots \\
& - CmpVal_{zz} \times Total_{zz} \\
& - (GapCreationPenalty \times GapNumber) \\
& - (GapExtensionPenalty \times TotalLengthOfGaps)
\end{aligned}
$$

For a more complete discussion of scoring matrices, see the **Data Files** manual.

CONSIDERATIONS

BestFit Always Finds Something

BestFit always finds an alignment for any two sequences you compare -- even if there is no significant similarity between them! You must evaluate the results critically to decide if the segment shown is not just a random region of relative similarity.

The Segments Shown Obscure Alternative Segments

BestFit only shows one segment of similarity, so if there are several, all but one is obscured. You can approach this problem with graphic matrix analysis (see the Compare and DotPlot programs). Alternatively, you can run BestFit on ranges outside the ranges of similarity found in earlier runs to bring other segments out of the shadow of the best segment.

The Best Fit is Only One Member of a Family

Like all fast gapping algorithms, the alignment displayed is a member of the family of best alignments. This family may have other members of equal quality, but will not have any member with a higher quality. The family is usually significantly different for different choices of gap creation and gap extension penalties. See the CONSIDERATIONS topic in the entry for the Gap program in the **Program Manual** to learn more about how to assign gap creation and gap extension penalties.

The Surface of Comparison

The magnitude of the computer's job is proportional to the area of the surface of comparison. That area is determined by the product of the lengths of the two sequences compared. BestFit can evaluate a surface of up to 3.5 million elements. This surface would be large enough to compare two sequences approximately 1,870-symbols long, or one sequence 200-symbols long with another sequence 17,500-symbols long. When you have much longer sequences that are known to align well, you can use the command-line option `-LIMit` to use the surface more efficiently.

The Public Scoring Matrix for Nucleic Acid Comparisons is Very Stringent

The scoring matrix swgapdna.cmp penalizes mismatches -0.9 so the segments found may be very brief. This penalty means that the alignment cannot be extended by three bases to pick one extra match. The scoring matrix used by Smith and Waterman, when local alignments were first described, used -0.333 for the mismatch penalty. You can use Fetch to copy randomdna.cmp and rename it swgapdna.cmp to use these values, or use nwsgapdna.cmp, which has no mismatch penalty at all.

Rapid Alignment

When possible, BestFit tries to find the optimal alignment very quickly. If this rapid alignment is not unambiguously optimal, BestFit automatically realigns the sequences to calculate the optimal alignment. When this occurs, the monitor of alignment progress on your terminal screen (`Aligning...`) is displayed twice for a single alignment.

ALIGNING LONG SEQUENCES

This program can align very long sequences if you know roughly where the alignment of interest begins. Run the program with the command line option `-LIMit`. Then set the starting coordinates for each sequence near the point where the alignment of interest begins and set gap shift limits on each sequence. The program then aligns the sequences from your starting point such that the sequences do not get out of phase by more than the gap shift limits you have set. If you started both sequences at

base number one and set the gap shift limit for sequence one to 100 and for sequence two to 50, then base 350 in sequence one could not be gapped to any base outside of the range from 300 to 450 on sequence two.

If you omit `-LIMit` on the command line, the program automatically sets gap shift limits if they are needed to allow the alignment of long sequences to proceed. In this case, the program limits the total length of gaps that can be inserted into each sequence and calculates the best alignment within this incomplete, or *limited*, surface of comparison. The program then performs a calculation to determine whether the alignment could possibly be improved if there were no restriction on the total length of gaps in each sequence. If the program cannot rule out this possibility, it displays the message `*** Alignment is not guaranteed to be optimal ***`. Because the criteria used in the calculation for guaranteeing an optimal alignment are very stringent, a limited alignment often may be optimal even if this message is displayed. In any event, the program continues to completion.

EVALUATING ALIGNMENT SIGNIFICANCE

This program can help you evaluate the significance of the alignment, using a simple statistical method, with the `-RANdomizations` command line option. The second sequence is repeatedly shuffled, maintaining its length and composition, and then realigned to the first sequence. The average alignment score, plus or minus the standard deviation, of all randomized alignments is reported in the output file. You can compare this average *quality* score to the quality score of the actual alignment to help evaluate the significance of the alignment. The number of randomizations can be specified along with the `-RANdomizations` command line qualifier; the default is 10.

The score of each randomized alignment is reported to the screen. You can use `<Ctrl>C` to interrupt the randomizations and output the results from those randomized alignments that have been completed.

By ignoring the statistical properties of biological sequences, this simple Monte Carlo statistical method may give misleading results. Please see Lipman, D.J, Wilbur, W.J., Smith, T.F., and Waterman, M.S. (Nucl. Acids Res. 12; 215-226 (1984)) for a discussion of the statistical significance of nucleic acid similarities.

ALIGNMENT METRICS

BestFit and Gap display four figures of merit for alignments: Quality, Ratio, Identity, and Similarity.

The Quality (described above) is the metric maximized in order to align the sequences. Ratio is the quality divided by the number of bases in the shorter segment. Percent Identity is the percent of the symbols that actually match. Percent Similarity is the percent of the symbols that are similar. Symbols that are across from gaps are ignored. A similarity is scored when the scoring matrix value for a pair of symbols is greater than or equal to 0.50, the *similarity threshold.* This threshold is also used by the display procedure to decide when to put a ':' (colon) between two aligned symbols. You can reset it from the command line with the second optional parameter of `-PAIr`. For instance, the expression `-PAIr=1.0,0.5` would set the similarity threshold to 0.5.

The similarity and identity metrics are not optimized by alignment programs so they should not be used to compare alignments.

PEPTIDE SEQUENCES

If your input sequences are peptide sequences, this program uses a scoring matrix with matches scored as 1.5 and mismatches scored according to the evolutionary distance between the amino acids as measured by Dayhoff and normalized by Gribskov (Gribskov and Burgess Nucl. Acids Res. 14(16); 6745-6763 (1986)).

RESTRICTIONS

Input sequences may not be more than 30,000-symbols long. This program cannot evaluate a surface of comparison larger than 5.5 million elements. A 200 x 27,500 comparison is possible, as well as a 2,300 x 2,300 comparison. See the ALIGNING LONG SEQUENCES topic for help in aligning long sequences that would normally exceed the maximum surface of comparison. You can also ask your system manager to increase the maximum surface of comparison if your system has enough virtual memory.

SEQUENCE TYPE

The function of BestFit depends on whether your input sequence(s) are protein or nucleotide. Normally the type of a sequence is determined by the presence of either `Type: N` or `Type: P` on the last line of the text heading just above the sequence itself. If your sequence(s) are not the correct type, turn to Appendix VI for information on how to change or set the type of a sequence.

COMMAND-LINE SUMMARY

All parameters for this program may be put on the command line. Use the option `-CHEck` to see the summary below and to have a chance to add things to the command line before the program executes. In the summary below, the capitalized letters in the qualifier names are the letters that you *must* type in order to use the parameter. Square brackets ([and]) enclose qualifiers or parameter values that are optional. For more information, see "Using Program Parameters" in Chapter 3, Basic Concepts: Using Programs in the **User's Guide.**

```
Minimal Syntax: % bestfit [-INfile1=]gamma.seq  [-INfile2=]alu.seq -Default

Prompted Parameters:

-BEGin1=1 -BEGin2=1       beginning of each sequence
-END1=500 -END2=207       end of each sequence
-NOREV1     -NOREV2       strand of each sequence
-GAPweight=5.0            gap creation penalty     (3.0 is protein default)
-LENgthweight=0.3         gap extension penalty    (0.1 is protein default)
[-OUTfile1=]gamma.pair    output file for alignment

Local Data Files: -DATa=swgapdna.cmp  scoring matrix for nucleic acids
                  -DATa=swgappep.cmp  scoring matrix for peptides

Optional Parameters:

-OUTfile2=gamma.gap       new sequence file for sequence 1 with gaps added
-OUTfile3=alu.gap          "     "       "   "      "    2   "   "     "
-LIMit1=499 -LIMit2=206 limit the surface of comparison
-RANdomizations[=10]      determine average score from 10 randomized
                              alignments
-PAIr=1.0,0.5,0.1         thresholds for displaying '|', ':', and '.'
-WIDth=50                 the number of sequence symbols per line
-PAGe=60                  adds a line with a form feed every 60 lines
-NOBIGGaps                suppresses abbreviation of large gaps with '.'s
-HIGhroad                 makes the top alignment for your parameters
-LOWroad                  makes the bottom alignment for your parameters
-NOSUMmary                suppresses the screen summary
```

ACKNOWLEDGEMENTS

Gap and BestFit were originally written for Version 1.0 by Paul Haeberli from a careful reading of the Needleman and Wunsch (J. Mol. Biol. **48**; 443-453 (1970)) and the Smith and Waterman (Adv. Appl. Math. **2**; 482-489 (1981)) papers.

Limited alignments were designed by Paul Haeberli and added to the Package for Version 3.0. They were united into a single program by Philip Delaquess for Version 4.0. Default gap penalties for protein alignments were modified according to the suggestions of Rechid, Vingron and Argos (CABIOS **5**; 107-113 (1989)).

LOCAL DATA FILES

The files described below supply auxiliary data to this program. The program automatically reads them from a public data directory unless you either 1) have a data file with exactly the same name in your current working directory; or 2) name a file on the command line with an expression like `-DATa1=myfile.dat`. For more information see Chapter 4, Using Data Files in the **User's Guide**.

If the first sequence you name is a nucleic acid, BestFit uses the scoring matrix in the public file swgapdna.cmp. (SW stands for Smith and Waterman.) If the first sequence you name is a peptide sequence, BestFit reads swgappep.cmp instead. The presence of these files in your current working directory causes BestFit to read your version instead. (See the **Data Files** manual for more information about scoring matrices.)

OPTIONAL PARAMETERS

The parameters and switches listed below can be set from the command line. For more information, see "Using Program Parameters" in Chapter 3, Basic Concepts: Using Programs in the **User's Guide**.

`-LIMit1=20` and `-LIMit2=20`

let you set *gap shift limits* for each sequence. When you already know of a long similarity between two sequences you can "zip" them together using this mode. The beginning coordinates for each sequence must be near the beginning of the alignment you want to see. The alignment continues so that gaps inserted do not require the sequences to get out of step by more than the gap shift limits. You can align very long sequences rapidly. The surface of comparison is still limited to 3.5 million. The size of a comparison can be predicted by multiplying the average length of the two sequences by the sum of the two shift limits.

If you add `-LIMit` to the command line without any qualifier value, the program prompts you to enter gap shift limits for each sequence.

`-RANdomizations=10`

reports the average alignment score and standard deviation from 10 randomized alignments in which the second sequence is repeatedly shuffled, maintaining the length and composition of the original sequence, and then aligned to the first sequence. You can use the optional parameter to set the number of randomized alignment to some number other than 10.

`-OUTfile2=seqname1.gap -OUTfile3=seqname2.gap`

This program can write three different output files. The first displays the alignment of sequence one with sequence two. The second is a new sequence file for sequence one, possibly expanded by gaps to make it align with sequence two. The third, like the second, is a new sequence file for sequence two, possibly expanded by gaps to make it align with sequence one. The program writes only the first file unless there are output file options on the command line. If there are any output files named on the command line, *only* those output files are written. If you add

-OUT to the command line without any qualifying filename, then the program will write the second and third output files after prompting you for their names.

Aligned sequences (in sequence files) can be displayed with GapShow. Their similarity can be displayed with PlotSimilarity.

-PAIr=1.0,0.5,0.1

The paired output file from this program displays sequence similarity by printing one of three characters between similar sequence symbols: a pipe character(|), a colon (:), or a period (.). Normally a pipe character is put between symbols that are the same, a colon is put between symbols whose comparison value is greater than or equal to 0.50, and a period is put between symbols whose comparison value is greater than or equal to 0.10. You can change these *match display thresholds* from the command line. The three parameters for -PAIr are the display thresholds for the pipe character, colon, and period. The match display criterion for a pipe character changes from symbolic identity (the default) to the quantitative threshold you have set in the first parameter. A pipe character will no longer be inserted between identical symbols unless their comparison values are greater than or equal to this threshold. If you still want a pipe character to connect identical symbols, use x instead of a number as the first parameter. (See the **Data Files** manual for more information about scoring matrices.)

-PAGe=64

When you print the output from this program, it may cross from one page to another in a frustrating way -- especially when you print on individual sheets. This option adds form feeds to the output file in order to try to keep clusters of related information together. You can set the number of lines per page by supplying a number after the -PAGe qualifier.

-WIDth=50

puts 50 sequence symbols on each line of the output file. You can set the width to anything from 10 to 150 symbols.

-NOBIGGaps

suppresses large gap abbreviations, showing all the sequence characters across from large gaps. Usually, gaps that extend one sequence by more than one complete line of output are abbreviated with three dots arranged in a vertical line.

-LOWroad and -HIGhroad

The insertion of gaps is, in many cases, arbitrary, and equally optimal alignments can be generated by inserting gaps differently. When equally optimal alignments are possible, this program can insert the gaps differently if you select either the -LOWroad or the -HIGhroad options. Here are examples for the alignment of GACCAT with GACAT with different parameters.

```
For:        Match = 1.0        MisMatch = -0.9
      Gap weight = 1.0   Length Weight =  0.0

LowRoad:   1 GACCAT 6
             .  |||      Quality = 4.0
           1 GA.CAT 5

HighRoad:  1 GACCAT 6
             ||| ||      Quality = 4.0
           1 GAC.AT 5
```

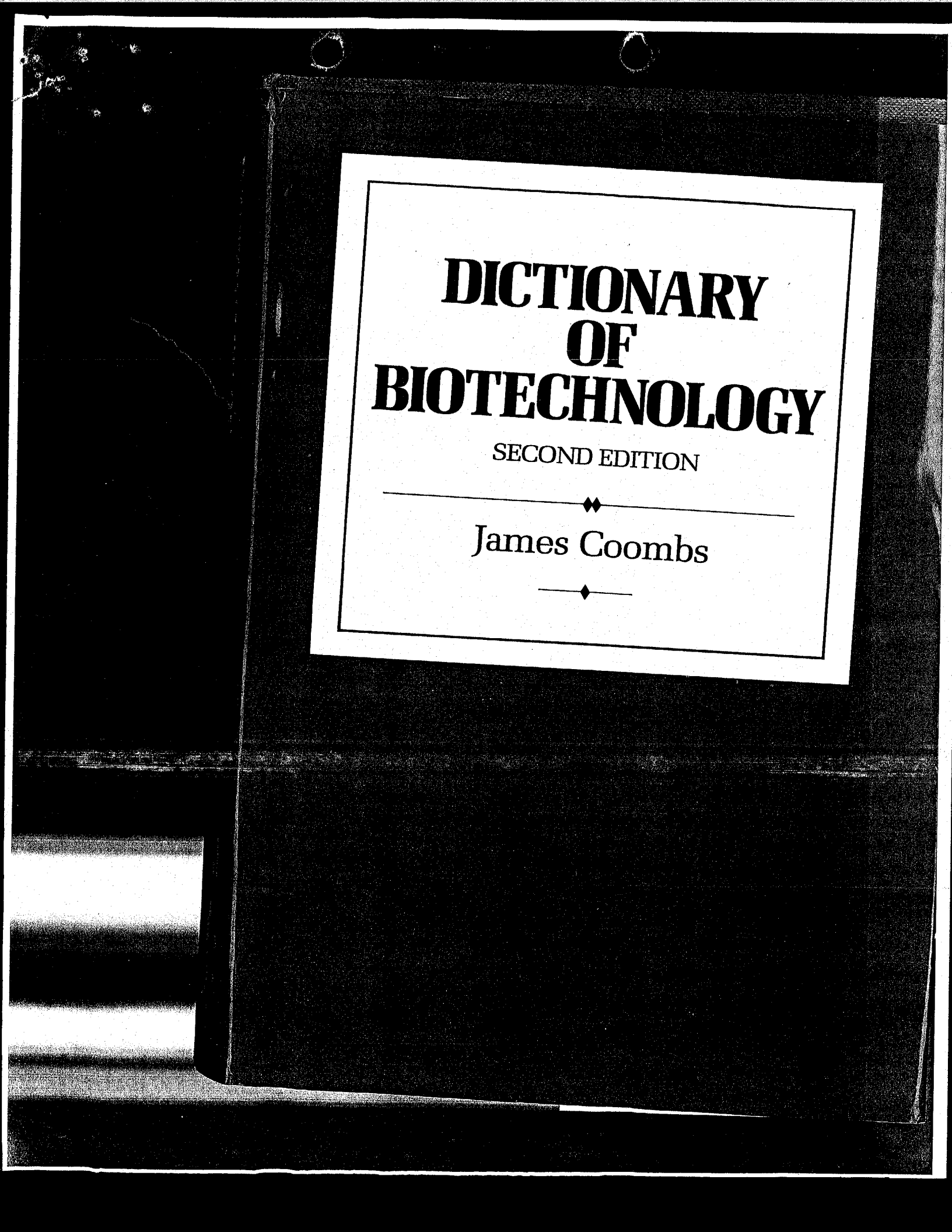
```
For:        Match = 1.0        MisMatch =  0.0
      Gap weight = 3.0   Length Weight =  0.0

HighRoad:  1 GACCAT 6
             |||           Quality = 3.0
           1 GACAT. 5

LowRoad:   1 GACCAT 6
             ..|||         Quality = 3.0
           1 .GACAT 5
```

Essentially the *low road* shifts all of the arbitrary gaps in sequence two to the left and all of the arbitrary gaps in sequence one to the right. The *high road* does exactly the opposite. When neither *high road* nor *low road* is selected, the program tries not to insert a gap whenever that is possible and uses the high road alternative for all collisions.

`-SUMmary`

writes a summary of the program's work to the screen when you've used the `-Default` qualifier to suppress all program interaction. A summary typically displays at the end of a program run interactively. You can suppress the summary for a program run interactively with `-NOSUMmary`.

Use this qualifier also to include a summary of the program's work in the log file for a program run in batch.


DICTIONARY
OF
BIOTECHNOLOGY
SECOND EDITION
James Coombs

Published in the United States and Canada by
STOCKTON PRESS, 1992
257 Park Avenue South,
New York, N.Y. 10010, USA

ISBN 1-56159-074-6

Second edition first published 1992 by
THE MACMILLAN PRESS LTD
London and Basingstoke

Associated companies in Auckland, Delhi, Dublin, Gaborone, Hamburg, Harare, Hong Kong, Johannesburg, Kuala Lumpur, Lagos, Manzini, Melbourne, Mexico City, Nairobi, New York, Singapore, Tokyo.

A catalogue record for this book is available from The British Library.

ISBN 0-333-57822-8

Printed in Great Britain.

tions that enables antibodies to be detected in the presence of known antigens and vice versa.

complementary bases Pairs of bases (purines and pyrimidines) that associate through hydrogen bonding in double-stranded nucleic acid. The following base pairs are complementary: guanine and cytosine; adenine and thymine, adenine and uracil.

complementary DNA *See* cDNA.

complementary sequences Two sequences of nucleotides that are capable of base pairing throughout their length.

complementary strands Two single strands of DNA in which the nucleotide sequence is such that they will bind as a result of base pairing throughout their full length.

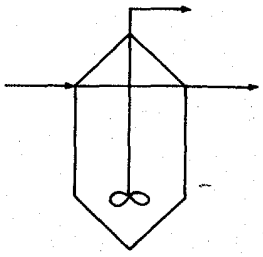
complementation The situation in which a normal wild-type phenotype is formed when two homologous chromosomes, known to bear a mutant gene, are brought together in a diploid cell.

complementation test A genetic test to ascertain whether two gene mutations occur in the same functional gene and to establish the limits of the functional gene; also called the *cis–trans* test.

completely mixed bioreactor A stirred tank fermenter or a continuous single-stage, once through microbial process that does not involve recycle of the cell biomass. The contents of the bioreactor are mixed by intermittent or continuous stirring of the liquid. This is achieved by mechanical means (using an impeller) or by recycling liquid or passing a gas through it.

complexity The number of units in a non-repeating sequence of nucleotide pairs in a prokaryotic genome or a haploid complement of chromosomes.

composting A process used to hasten the aerobic decomposition of organic wastes (horticultural, agricultural or municipal)



Completely mixed continuous digester

resulting in the production of a humus rich soil. The decay process is the result of the combined action of invertebrates, including insects and worms, as well as bacteria and fungi. Industrial composting technologies include various techniques for aerating the compost, as well as the addition of cultures of bacteria (including nitrogen fixers), cellulolytic or lignolytic fungi, or earthworms in order to increase the rate of decomposition.

compound A substance composed of one type of molecule only.

computer integrated fermentation (CIF) A real-time process data management system that integrates itself and computer assisted software tools into the overall fermentation process. CIF can control several fermenters simultaneously, allowing data exchange between different processes.

concanavilin A A lectin, isolated from the legume *Canavalia ensiformis*, that binds specifically to the glucose and mannose residues on the surface of transformed cells, producing agglutination.

concatemer A DNA structure made up of linearly repeated unit length DNA molecules.

concatemeric Descriptive of DNA molecules or sequences (not necessarily identical) covalently linked in series.

concentration The amount of a particular compound in a defined volume.

concentration gradient A solution in which the ratio of solvent to solute changes

GBREL.TXT Genetic Sequence Data Bank
15 December 1999

NCBI-GenBank Flat File Release 115.0

Distribution Release Notes

5354511 loci, 4653932745 bases, from 5354511 reported sequences

This document describes the format and content of the flat files that comprise releases of the GenBank database. If you have any questions or comments about GenBank or this document, please contact NCBI via email at info@ncbi.nlm.nih.gov or:

GenBank
National Center for Biotechnology Information
National Library of Medicine, 38A, 8N805
8600 Rockville Pike
Bethesda, MD 20894
USA
Phone: (301) 496-2475
Fax: (301) 480-9241

TABLE OF CONTENTS